DOCUMENT RESUME

ED 183 618 TM 800 031

AUTHOR Surkan, Alvin J.; Evans, Richard M.
TITLE Interactive Testing in Introductory Educational Measurement.
PUB DATE 11 Apr 79
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 9-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Programs; Higher Education; Item Analysis; Measurement; *Measurement Techniques; *Test Interpretation
IDENTIFIERS *Computer Assisted Testing

ABSTRACT

Since fall 1976, an undergraduate measurement class has utilized a 32K microcomputer programmed in APL to present replicate sets of questions randomly selected from 80-question item pools representing each of three domains of knowledge based on Gax's Principles of Educational Measurement and Evaluation. Multiple choice questions are presented on a cathode ray tube and students are required to answer until a correct choice has been made. Feedback frames, concerning the rightness and wrongness of the question's alternatives are next presented. After 15 such iterations, grades are calculated with criteria of 85, 75, and 65 percent, respectively, for grades of A, B, or C. This paper describes the system's functions, namely: (1) test creation utilities, (2) the interactive test control and question management functions, and (3) test item analysis functions. Some history and research in using the system, general empirical observations, and a discussion of some implications of the system in creating new possibilities for reforming the basis of education are presented. (Author/JKS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED183618

TM800 031

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Interactive Testing in Introductory Educational Measurement

Alvin J. Surkan

Department of Computer Science

University of Nebraska - Lincoln 68588

and

Richard M. Evans

Department of Education

Morningside College - Sioux City, Iowa 51106

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard M. Evans

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the annual meeting of the National Council on Measurement in Education, Session 14B, 11 April 1979.

INTERACTIVE TESTING IN INTRODUCTORY EDUCATIONAL MEASUREMENT

Introduction

Since Fall Semester 1976, an undergraduate measurement class has utilized a lowcost 32K microcomputer programmed in APL to present replicate sets of questions randomly selected from 80-question item pools representing each of three domains of knowledge within the course. The domains are represented by chapters 1-5, 6, and 7-12 of Sax's Principles of Educational Measurement and Evaluation, (1974). Multiple choice questions were presented on a cathode ray tube and students were required to answer until a correct choice had been made. Feedback frames, concerning the rightness and wrongness of the question's alternatives were next presented. After 15 such iterations grades were calculated with criteria of 85, 75, and 65 percent, respectively, for grades of A, B, or C.

This system has been used in every section of the course following that introduction--a total of four times involving a total of 79 students. The system was originally introduced in an educational psychology class during the Spring Semester 1976, and has been tested with two other courses: Human Sexuality and Logic.

The system has allowed the students to control when they take their own examinations. It enables them to retake an exam in a domain of knowledge several times, if necessary, with no penalty for earlier transgressions. It provides the instructor with diagnostic information concerning an individual's difficulties with course material earlier in the semester than was possible before. It has effectively separated the teaching and testing functions from the act of instruction, making the teacher a helper and the tests themselves a learning experience.

This paper first describes the system's functions, namely: (1) test creation utilities, (2) the interactive test control and question management functions, (3) test item analysis functions. This is followed by: (a) some history and research using the system, (b) general ~~somewhat~~ empirical observations and (c) a discussion of some implications of the system in creating new possibilities for reforming the basis of education.

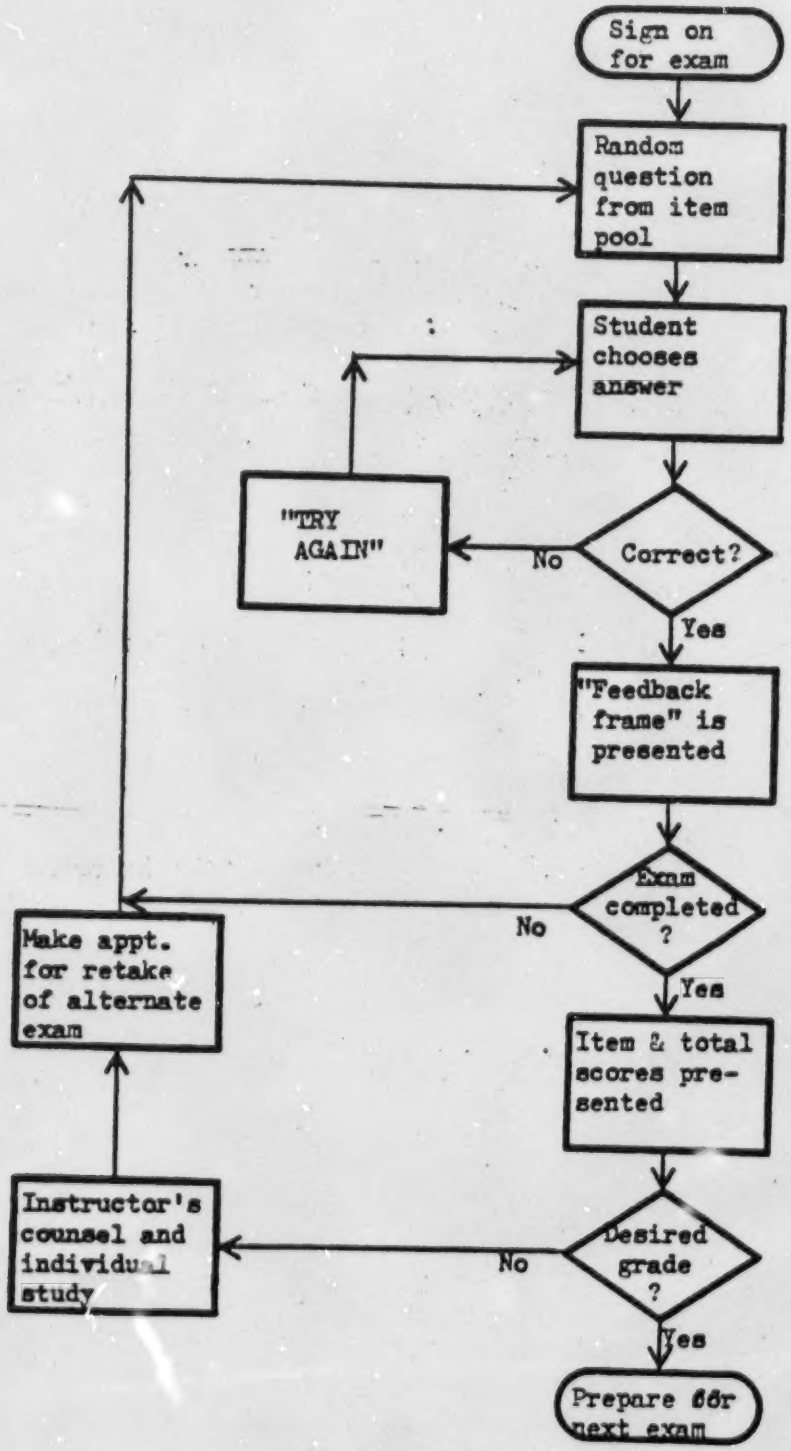

Figure 1. Mastery strategy for time-variable competency-based course of study.

Utility Functions for Test Creation

During the preparation of test items the following APL functions are used: (1) TEXTIN, (2) PUTQAF, (3) FIX, and (4) GETQAF. The TEXTIN function facilitates the entry of character strings which constitute the question (4-way multiple choice), the correct answer and the feedback or reinforcement. After TEXTIN is used to fill the QQ, AA and FF with question, answer and feedback, respectively, PUTQAF is used to store these components of the completed item in a tape file of the cartridge used for item storage. Either before using PUTQAF or any time after using GETQAF to retrieve a copy of an item from its tape file, the function FIX may be used to alter any line of characters in the item of current interest. Once this fixing is done, PUTQAF may be used again for sending the new version of item back to file storage on tape.

Functions for Control and Management of Testing

To reset a variable called IDNOS, which keeps track of the testing that has gone on, a function NEWCLASS is provided. This function can be run only by the person capable of producing a password requested by a locked function called PASSWORD. Upon receiving the proper password, NEWCLASS sets to zero IDNOS and instructs the user on how to remark the scorekeeping tape cartridge to purge and retag with TAGFILE all 107 files beyond the first four. Finally, instructions are given for admitting the students for a new class using the function provided with the name ADMIT.

Testing and Reinforcement Functions

The main function for entry by a student into a test is called STUDENT. This function sequences three activities: (1) the identification of student by name and obtaining the date and test numbers using a function IDT, (2) the checking and requesting of correct tape identification by the SEEK function, and (3) the actual conduct of the testing and reinforcement by function QUIZ.

The IDT function first uses SEEK to establish that the grade tape is in place and recovers from the first file the IDNOS used for controlling the testing of students admitted to the course IDT also.

The QUIZ function generates and orders a preselected set of random numbers for selecting items. Sequentially, these items consisting of questions, answers and feedback are presented and the student is assigned a partial score for each based on how few attempts are required to give the correct answer. During the item presentation a function DELAY is frequently used to insure that the information such as instructions, questions, reinforcement, and scores are kept on the CRT screen for proper amounts of time for reading. Upon completion of item presentation the QUIZ function shows the scores earned, requests the grade tape cartridge to be installed, and insures that student individual score files are updated.

Functions for Score Display and Recording

During the conduct of a test by QUIZ a function called STATISTICS collects the score for each question presented in a variable called SCORE. At the end of a testing session a function called SHOW tabulates and puts the RESULTS produced by function FORMAT operating on SCORE under a heading. The SHOW function also displays the test number, the total points earned by the student and his grade as a percentage along with the total possible points for the number of items presented. The function SHOW may also be used by the student after completion of his test to display the table of results upon request. Just before completion of a test session, a function SEEK is used to establish or request that the cartridge with scores is in place. Subsequent to this, a function called APPEND is used by QUIZ to update the student's scores on tape. This is effected by evoking a function called FETCH which retrieves previously packed scores for the students file. APPEND then uses PAC to compact the new scores and append them to the previous scores; then the combined set of scores are all sent back out to the student's tape file.

Test Item Analyses Functions

Functions for test item analyses perform selective retrieval of subsets of scores from the grade files organized by the student and place them in item-based files with data on items aggregated over specified subsets of student interaction. The item oriented files of data can be used for evaluating variables estimators proposed for determining the relative performance of each item.

Functions used for selective recovery and item scores include FETCH, UPC, SELECT, SELECTOR MEMBER, HEAD. Functions for computing statistical estimators of the items are DIFF, DISC, USE, STATS. The functions of the RETRIEVE work spaces are RETRIEVE FILTER, UPC, MBR, SCRRSTAT, HEAD, ITEM, SETUP.

Measures Used in Item Analysis

Data collected on the responses of students to individual teaching are used for computing item performance measures over each item's use. The three computed quantities tabulated are (1) difficulty, (2) discrimination, and (3) use. Each quantity is tabulated for the three trials individually by item as shown in figure 2.

Four measures of performance for each complete item pool sampled for testing students have been computed over all items and use are computed for monitoring, as a function of trial, the behavior of aggregate quantities (1) mean, (2) standard deviation, (3) KR20, and (4) standard error of measurement.

Kuder-Richardson Measure of Test Reliability

$$KR_{20} = \frac{N}{N-1} * \left(\frac{V - \sum_{I=1}^{N} PQ}{V} \right)$$

where

N = number of different test items contributing to the reliability estimate for each trial

P = estimate of average difficulty for each distinct test item within a trial

Q = complement of difficulty or 1-P

V = variance of all scores of students having used the N distinct test items within a trial

Standard Error of Measurement Estimate for Test Scores

$$E = S * (1 - KR_{20})^{1/2}$$

where

S = standard deviation of test scores within a trial

KR_{20} = Kuder-Richardson reliability measure for all test scores within a trial

Standard Deviation of Test Scores

$$S = \left\{ \frac{1}{N} \sum (Y - \bar{Y})^2 \right\}^{1/2}$$

where

Y = values of test scores within a trial

$\bar{Y}$ = mean of test scores within a trial

N = number of given test scores within a trial

Item Performance Measure

Defined discrimination measure

$$D = \sum_{i=1}^{M} \left(\frac{Y-\bar{Y}}{\bar{Y}}\right) \cdot \left(\frac{X-\bar{X}}{\bar{X}}\right)$$

where:

X = partial score on item by student

Y = total score on N item test trial

M = times the item was used in given trial category

N = number of items on the test

P = number of points per item

I = index on item use

Since the distribution of X and Y are not known one can get first order approximation using

$\bar{X}$ = 1/1 P

$\bar{Y}$ = 1/2 NP

yielding

$$D = \frac{4}{M} \sum_{i=1}^{M} \left(\frac{Y}{NP} - \frac{1}{2}\right) \cdot \left(\frac{X}{3} - \frac{1}{2}\right)$$

for N = 15 and P = 3

$$D = \frac{4}{M} \sum_{I=1}^{M} \left(\frac{Y}{45} - \frac{1}{2}\right) \cdot \left(\frac{X}{3} - \frac{1}{2}\right)$$

The contributions under the summation sign can be accumulated along with N during each testing session. At the end of each session they can be used to update the performance data for those items that appeared in the sample comprising the session's test.

[illegible] 4 ITEMSTUDY

A 12
RESULTS

MEANS [illegible]
STDVS [illegible]
KR20S [illegible]
STDKS [illegible]

STABLE

QUESTION	TRIAL NO. 1			TRIAL NO. 2			TRIAL NO. 3		
	DIFF	DISC	USE	DIFF	DISC	USE	DIFF	DISC	USE
1	100	54	3	100	87	1	100	62	2
2	100	73	2	67	10	1	0	-29	1
3	33	6	2	100	56	1	83	20	2
4	56	3	3	42	11	4	100	78	1
5	100	54	6						
6	33	-11	4						
7	89	34	3				67	34	2
8	94	40	6	83	35	2	100	29	1
9	50	-1	4	83	48	4			
10	100	53	2	50	22	2	33	-16	1
11	33	-22	2	67	27	1	100	78	1
12	75	11	4	89	47	3	100	29	1
13	13	-36	5	56	21	3			
14	89	42	3	50	9	2	100	47	1
15	100	64	2	100	71	2			
16	100	33	3	93	61	5			
17	80	36	5	67	20	5	67	10	1
18	67	18	2	0	-29	1	0	-47	1
19	75	38	4						
20	67	10	1	100	60	2			
21	67	31	4	100	78	2	67	10	1
22	67	21	1	67	19	1			
23	78	37	3	100	29	1	100	78	1
24				89	28	3			
25	67	37	2	89	55	3			
26	33	-7	5	50	18	2			
27	50	18	4	100	82	1	100	47	1
28	56	-1	3	0	-29	1	83	36	2
29	100	53	2				67	10	1
30	67	30	2	100	67	2	33	-21	2
31	78	31	6	100	51	3	100	47	1
32	50	14	4	83	46	2	100	78	1
33	89	37	3	92	39	4			
34				83	34	4	100	29	1
35	67	40	2	67	10	2			
36	67	10	4	33	-10	1	0	-47	1
37	67	37	4						
38	100	58	2	33	-16	1	100	78	1
39	83	44	4	100	29	1	100	47	1
40	100	35	3	100	53	3			
41	100	11	1	100	56	1			
42	0	-24	1	83	44	2			
43	53	12	5	83	-48	2			
44	60	-2	5	33	-11	1	100	29	1
45	100	54	4	89	45	3			
46	100	51	1	92	45	4	100	47	1
47	33	-8	1	83	21	2			
48	92	29	4	83	46	4	100	78	1
49	33	-8	1	92	57	4	67	10	1
50	67	37	4	67	33	3			
51	78	16	3	67	18	2			
52	100	67	2	100	60	5	100	78	1
53	50	-9	2	78	22	3	100	47	1
54	78	47	3				33	-16	1
55	78	19	3	71	23	5			
56	58	27	4	100	56	1			
57	83	39	2				67	26	1
58	100	11	1	50	-2	2	67	19	2
59	100	54	4	100	72	4			
60	100	38	2				33	-10	1
61	81	25	7	67	25	4			
62	33	8	2						
63	50	27	2	100	78	1	100	78	1
64	71	26	7	33	-11	1			
65	94	53	6	89	60	3			
66	83	24	2				33	-10	1
67	56	19	3	0	-42	2	83	27	2
68	83	24	4				100	70	1

TO SEE THIS STATISTICS TABLE AGAIN, ENTER: STABLE

Figure 2. Item analysis data for Test 2

History and Research of the System

The project began in 1973 when "feedback forms" were developed as answer sheets for five multiple-choice examinations using the Biehler (1974) educational psychology materials. The answer sheets used the A.B. Dick "latent image" invisible ink process which allowed students to use an answer-until-correct (AUC) procedure when taking examinations (Hanna, 1977). The AUC procedure deals with Skinner's (1968) criticism of multiple-choice tests by not allowing students to exit the questions until they have made a correct choice. An additional advantage of AUC scoring over right-wrong scoring is increased split-half test reliabilities (Evans and Misfeldt, 1974). Following completion of the 35 question examinations, each student was given a booklet explaining why three of the four answers of each multiple-choice questions were wrong, and why the remaining answer was deemed "correct".

Salient aspects of this system were integrated through use of a dependable, ~~reliable~~, dedicated microprocessor-based computer that became commercially available late in 1975. An integrated interactive testing, scoring, and recordkeeping system was developed in APL (an array oriented interpreter-based language) in an experiment with the educational psychology class of the 1976 Spring Semester. The same Biehler materials, with five examinations, had been used in the course the previous year without a computer. Comparing subsequent class results with those from previous Spring Semesters revealed that students using the interactive system did more retakes for all of the five examinations, scored higher on three of the five examinations and had a stronger perception of learning more with the computer testing than did a similar group who studied a year earlier (Evans and Surkan, 1977a). The known differences between the two groups were in (1) method of test administration and response recording, and (2) the time of five class periods that would normally be used for traditional classroom testing which was gained for instruction since all trials on the examinations were made outside of class.

Important similarities between the groups were (1) the instructor, course materials, and knowledge domains for sets of test items were the same, (2) the AUC procedure was retained, (3) ACT scores, and (4) explanations of test items were with the help of the computer presented immediately after obtaining each correct response to each of 15 items in the test sample, rather than postponed until after the completion of a 35-question examination.

Since out-of-class testing allowed students control over the time at which each examination was taken, it soon became interesting to examine the variation in time among students as they progressed through such time-variable, competency-based (TVCB) courses of study. By this time the system had been applied to four courses of study: educational psychology, educational measurement, human sexuality, and logic. "Total days since the semester began" (TDAYS) was chosen as the dependent variable. An initial study found that time is an important variable in an instructional system that holds mastery level relatively constant (Evans and Surkan, 1977b), and can account for 18 percent of the final course grade variance. TDAYS also correlated negatively with achievement ($r = -.498$).

A factor analysis of another semester's educational psychology class also revealed that students who aspired to the higher levels of course mastery required less time (Evans and Surkan, 1978). This analysis also found that ACT scores were not correlated with final grades ($r = 0.00$), and that the so-called personality factors of "toughness" and "perseverance", had the most promise for serving as predictors when the TVCB aspects of a mastery learning instructional strategy diminishes the predictive value of academic aptitude measures. Thus, if it is true that almost all can learn, as proposed by Bloom (in Block, 1971) then variation in achievement is less interesting as a dependent variable than is time to mastery.

A study of the Fall 1976 educational psychology class found a positive

relationship between number of retakes of examinations and achievement level (Evans and Duncan, 1979). It was also found that student perceptions of the quality of study in preparation for examinations 4 and 5 related positively to achievement, and that hours of study in preparation for examinations related inversely with TDAYS on examinations 3, 4, and 5.

Scheduling up to 80 students in one or two courses during a given semester sometimes led to problems, since all were tested using a single microcomputer. It became hectic during the last few weeks of a semester, when the TVCB aspects of the courses became incompatible with severe external pressure from the registrar's imperative for grade submission. To minimize the end-of-semester queueing, due to earlier procrastination, grade criteria were modified to include incentives for early completion. The 1978 Spring Semester's education psychology class was rewarded for early completion by reducing the required percentage levels of the examination grade criterion by five percent if the test was taken before its subject material had been discussed in class, and raised by five percent, if the test was taken three or more weeks later than class completion of the complete topic and covered by the examination. A study of this effect found these students completing all five examinations in significantly fewer days than control groups from Spring Semester of the previous two years, while there were no significant differences among the groups in achievement scores (Evans and Johns, 1979).

Some General Observations

While the real life situations did not permit the foregoing research to be models of precision, experiment design, and rigor, three years of experience in four diverse courses of study enable those observing the system in operation on a daily basis builds confidence in the presented findings. Indeed, it is felt that the system is even more successful at testing-teaching-monitoring tasks than can be demonstrated by the currently available quantitative measurements. Accordingly, we will speculate about some non-systematic observations.

It appears that as efforts in providing feedback and formative evaluation are increased, more and more inadequacies in traditional instruction appear. Time spent counseling those individual students who fail to attain grade criterion on a particular attempt of an examination is at least quadrupled relative to the corresponding time with the previous mode of instruction. Student deficiencies are discovered earlier in the semester, when there is yet enough time to do something about them. This also expands the instructor's counseling time. Another reason for these increased office hours is the changed perception of the instructor's role from that of a judge, who must be kept from knowing about learning deficiencies, to that of a colleague who is helping the student in his mastering the knowledge in a neutral and untiring computer. The net effect is that the formal teaching and testing functions of instruction have been effectively separated although learning does continue during testing.

Before the item analysis model was completed, the instructor counseling sessions were used by the instructor for reviewing the questions the student had missed to determine which concepts needed remedial instruction. Upon discussing and examining of student thought processes bearing on an item, the instructor analysed the item itself for ambiguity, usefulness of this distractors,

and also attempted to defend the relavance of the item. This led to later revision of items. With the computer a new problem frequently arose that seldom occurred with paper-and-pencil testing. Specifically, students would often skim over the stem of a question superficially and then they would hastily choose a nice-sounding alternative, only to find they have made an error. Unfortunately, the nice-sounding alternative did not fit the stem--the student had failed to understand the stem thoroughly before moving down to the alternatives. Another reading problem that revealed itself during instructor counseling sessions concerns the random selection of 15 items from an 80 member item pool. On the second or third retake a student typically reencounters one or two items which were presented previously. This is acceptable because it encourages the student to read the feedback frames presented after every correct answer. It is perplexing that there exists a small number of students who may continue to miss questions they missed on previous examinations. Modification of the programs to provide extra drill for these students as they become so identified is being considered.

High student anxiety has been observed during the first few examinations. This seems to diminish as the semester progresses. A study is underway to assess changes in the IPAT (8 Parallel Forms Battery) during the present semester. A "systematic desensitization" of the test anxiety appears to be present because of the lack of dire consequences when criterion is not attained in an examination. Concommitant with this, there appears to be an increased aspiration in the mastery level sought in perhaps as much as 20 percent of a class. The "career B or C" students often begin to see themselves as "A" students. This is intrinsically rewarding to the instructor, particularly when it is observed in a student who has required several hours of out-of-class tutoring and counseling.

Some disturbing problems concerning the establishment of mastery criteria were discussed in the special issue of the Journal of Educational Measurement (NCME, 1978). The re-take option takes care of any inadvertent errors that may arise from setting arbitrary criteria. The measurement course began with the criteria of 90, 80, and 70 percent for grades of A, B, and C, respectively. These criteria used during the first semester of experimentation with the system proved to be too difficult, so the instructor tended to compensate by devoting more class time to motivation and increasing the out-of-class time allocated to diagnostic teaching. In ensuing semesters, these criteria were relaxed by five points for this course only.

Discussion

The computer based testing/teaching/student management system has proved successful in an undergraduate educational measurement course plus three other courses presented in a private liberal arts college. The system has introduced to students a degree of control over the time available for learning. This control seems to be related to increased student motivation and learning. It is concluded that such a system provides a conceptual framework for reform in school organization--from organization around similarly aged cohorts that move through a sequential curriculum at an average rate with varying levels of achievement to one comprising diverse individuals moving through a curriculum at their own rate to attain a preselected fixed level of mastery. The rapidly falling costs of microcomputer technology currently promises to make practical use of such a system for any subject matter amenable to behavioral definition. Thus, one may conceive of a school system organized more according to the nature of human learning rather than for the convenience of the people administering traditional systems.

Although the test item analysis functions were designed to use norm-referenced indices of item and test quality (in a domain criterion-referenced system), these functions offered the instructor a form of "formative evaluation" that was unlike that provided the student. One might ask: Why shouldn't the instructor also have constant feedback by monitoring the performance of the instruments in his system of instruction while they are in turn being used to monitor the individual student progress through segments of a knowledge domain? Such a monitoring system is analogous to the airline pilot's monitoring of instruments, autopilot, nagivational apparatus, and system warning lights while progressing through foul weather without reference to the horizon. Such flying is definitely <u>not</u> by the seat of one's pants. Likewise the educational process with modern instruction using a structured

curriculum must be monitored. When pilots observe a warning light indicating that one of the myriad of systems is malfunctioning in the aircraft, action must be taken to compensate or the flight must abort. The consequences of inaction in this analogy can be fatal. Instructors presently do not have timely feedback informing them about the operation of their instructional system while a course of study is in progress. Item analysis functions, when provided by adequately large computer systems, can provide such "warning lights". In aviation, the compensatory action taken after a warning of system malfunction is made in accordance with the "fail-safe" concept. This concept ensures that there are either back-up systems or there is human intervention in the case of failure of one of the systems. Perhaps this "Fail Safe" concept is highly appropriate for insturctional systems also.

Bibliography

Biehler, R. H. Psychology Applied to Teaching (2nd Ed.). Boston: Houghton Mifflin Company, 1974.

Bloom, B. S. Mastery Learning. In Block, J. H. Mastery Learning: Theory and Practice. New York: Holt, Rinehart and Winston, 1971.

Evans, R. M. and Duncan, D. A. Relationship of student perseverance, hours of study and quality of study, on time to mastery and level of achievement. Paper presented at American Educational Research Association Annual Meeting, San Francisco, April, 1979.

Evans, R. M. and Johns, J. C. Effect of criterion manipulation on time taken in self-paced learning. Proceedings, Association for Development of Computer-based Instructional Systems Annual Meeting, San Diego, California, February, 1979.

Evans, R. M. and Misfeldt, K. Effect of self-scoring procedures on test reliability. Perceptual and Motor Skills, 1974, 38, 1248.

Evans, R. M. and Surkan, A. J. Effect of computer-delivered testing on achievement in a mastery learning course of study with partial scoring and variable pacing. Proceedings, Eighth Conference on Computers in the Undergraduate Curricula, Michigan State University, 22 June 1977a. ERIC Document ED 156 160.

Evans, R. M. and Surkan, A. J. Time, ability, trials, and their relationship to achievement in a mastery learning instructional strategy. Mimeographed, 23 May 1977b.

Evans, R. M. and Surkan, A. J. Factors in time-variable learning allowed by computer delivered interactive testing. Proceedings, Association for Development of Computer-based Instructional Systems Annual Meeting Dallas, Texas, March 1978.

Hanna, G. S. A study of reliability and validity effects of total and partial immediate feedback in multiple-choice testing. Journal of Educational Measurement, 1977, 14, 1-7.

National Council of Measurement in Education. Journal of Educational Measurement, 1978, 15, 237-319.

Sax, G. Principles of Educational Measurement and Evaluation. Belmont, CA: Wadsworth Publishing Co., Inc. 1974.

Skinner, B. F. The Technology of Teaching, New York: Appleton-Century-Crofts, 1968.